International Journal of Modern Electronics and Communication Engineering (IJMECE) ISSN: 2321-2152
Volume No.-6, Issue No.-6, November, 2018

Modelling Physicochemical Properties for Protein Tertiary Structure Prediction: Performance Analysis of Regression Models

R.S. Kamath
Department of Computer Studies
Chhatrapati Shahu Institute of Business Education and Research
Kolhapur, India
E-mail: rskamath@siberindia.edu.in

R.K. Kamat
Department of Electronics
Shivaji University
Kolhapur, India
E-mail: rkk_eln@unishivaji.ac.in

Abstract: This paper explores performance analysis of various regression models for the prediction of protein tertiary structure by modelling physiochemical properties. The protein structure dataset for the study is retrieved from UCI Machine Learning repository. The research exhibits performance evaluation of various regression models and compares the prediction accuracy using R-squared value. The models include Decision Tree, Random Forest (RF), Neural Network and Linear Regression. The reported investigation depicts Random Forest model outperforms the rest of the models in prediction of protein tertiary structure. The measures of variable importance using RF algorithm reveals that physicochemical property F4 is stands at the top whereas F1 is least important.

Keywords: protein tertiary structure, machine learning, R data mining, random forest, biological significance

I. INTRODUCTION

Proteins are essential organic polymers shaped from building blocks called amino acids [2]. The three-dimensional structure and biological action of proteins rely upon the physicochemical properties of amino acids. The translation of protein sequences into tertiary forms is required to carry out various biological functions. The anticipation of high resolution protein structure is one of the great challenges in computational biology. Many research teams have carried out different models for the classification or prediction of protein structures.

Mishra et al have reported machine learning models for the classification of protein structures using physical and chemical properties [3]. The study concluded that random forest model is apt for the classification of protein structures. Gromiha has presented a simple linear regression model for the anticipation of protein folding rates using amino acid sequences [4]. This study has shown a good correlation between experimental and predicted values. Yet another paper by Pathak et al, explored machine learning models for the prediction of protein structures [5]. The result concluded that random forest outperforms the other model in structure prediction. The experiment conducted by Jani et al reported that Support Vector Machine with Principal Component Analysis feature extraction is performing better for the classification of protein tertiary structure [6]. Mishra and Ahiwar have reported the comparative study of supervised learning models for the prediction of protein structure [7]. Yet another research by Iraji and Ameri, proposed a soft computing model to reduce the predicted Root-mean-square-deviation error for protein tertiary structure [8]. Kamath and Kamat have depicted a decision tree model for protein expression levels for Down syndrome [9]. The experiment is simulated in R data mining environment.

In the backdrop of the research rendered here, this paper explores performance analysis of regression models for the prediction of protein tertiary structure. The dataset for the present study is taken from UCI repository [1]. The physicochemical properties of protein structure are analyzed using different prediction algorithms. Root Mean Square Deviation (RMSD) is an indicator of these algorithms. The performance accuracy reveals that Random Forest model is appropriate for predicting protein tertiary structure.

The rest of the paper is arranged as follows; brief introduction follows materials and methods in section two. The third section outlines computation details with results and discussion of machine learning techniques for protein tertiary structure prediction. The conclusion at the end explains the appropriateness of Random Forest model for modelling physiochemical properties of protein.

II. MATERIALS AND METHODS

The dataset comprises physicochemical properties of Protein Tertiary Structure is taken from CASP 5-9 [1]. The dataset contains 45730 instances based on nine features. These attributes are Total surface area, Non polar exposed area, Fractional area of exposed non polar residue, Fractional area of exposed non polar part of residue, Molecular mass weighted exposed area, Average deviation from standard exposed area of residue, Euclidian distance, Secondary structure penalty, Spacial Distribution constraints. RMSD is a response variable denotes root mean square deviation of protein tertiary structure. Exploratory analysis of the dataset is shown in figure 1.

```
       F1              F2              F3              F4              F5
Min.   : 2392   Min.   :  403.5   Min.   :0.09362   Min.   : 10.31   Min.   : 319490
1st Qu.: 6944   1st Qu.: 1980.2   1st Qu.:0.25898   1st Qu.: 63.54   1st Qu.: 955065
Median : 8908   Median : 2673.0   Median :0.29992   Median : 87.71   Median :1238137
Mean   : 9875   Mean   : 3018.6   Mean   :0.30253   Mean   :103.43   Mean   :1368759
3rd Qu.:12132   3rd Qu.: 3796.4   3rd Qu.:0.34285   3rd Qu.:133.15   3rd Qu.:1691596
Max.   :40035   Max.   :15312.0   Max.   :0.57769   Max.   :369.32   Max.   :5472011
       F6              F7              F8              F9             RMSD
Min.   : 31.97   Min.   :     0   Min.   :  0.00   Min.   :15.31   Min.   : 0.000
1st Qu.: 94.80   1st Qu.:  3166   1st Qu.: 31.00   1st Qu.:30.44   1st Qu.: 2.311
Median :126.42   Median :  3839   Median : 54.00   Median :35.31   Median : 5.110
Mean   :145.66   Mean   :  3985   Mean   : 69.94   Mean   :34.52   Mean   : 7.788
3rd Qu.:181.11   3rd Qu.:  4645   3rd Qu.: 91.00   3rd Qu.:38.86   3rd Qu.:13.450
Max.   :598.41   Max.   :105948   Max.   :350.00   Max.   :55.30   Max.   :20.999
```

Figure 1. Exploratory data analysis of protein tertiary structure dataset

The present study is carried out in Rattle, R data mining environment [10]. Dataset is partitioned randomly into training, testing and validation with division 70%, 15 % and 15% respectively. Following prediction algorithms are employed and the performance of the same is compared.

1. Decision Tree
2. Random Forest (RF)
3. Neural Network
4. Linear Regression

III. COMPUTATION DETAILS, RESULTS AND DISCUSSION

Figure 2 gives textual summary of regression tree derived in the present investigation. The package 'rpart' is used for the construction of this tree. It uses recursive partitioning [12]. That the root node of decision tree tests 'F3' value <= 0.33 continues down to the left side of the tree, otherwise right side of the tree. The next test down this right and left sides of the tree are F6 and F4 respectively. Thus, it proceeds and will be able retrieve RMSD value for selected instance. Thus derived regression tree is shown in figure 3.

```
Regression tree:
rpart(formula = RMSD ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)], method = "anova", parms = list(split = "information"),
    control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] F3 F4 F5 F6 F8

Root node error: 1200162/32010 = 37.493

n= 32010

        CP nsplit rel error  xerror      xstd
1 0.114855      0   1.00000 1.00002 0.0051260
2 0.031200      1   0.88515 0.88941 0.0053804
3 0.019828      2   0.85394 0.85873 0.0058601
4 0.017364      3   0.83402 0.83169 0.0058509
5 0.016534      4   0.81665 0.81779 0.0058632
6 0.016488      5   0.80012 0.81596 0.0058599
7 0.011906      6   0.78363 0.80127 0.0057742
8 0.010291      7   0.77172 0.78816 0.0058042
9 0.010000      8   0.76143 0.78275 0.0058195
```

Figure 2. Summary of regression tree using *"rpart"*

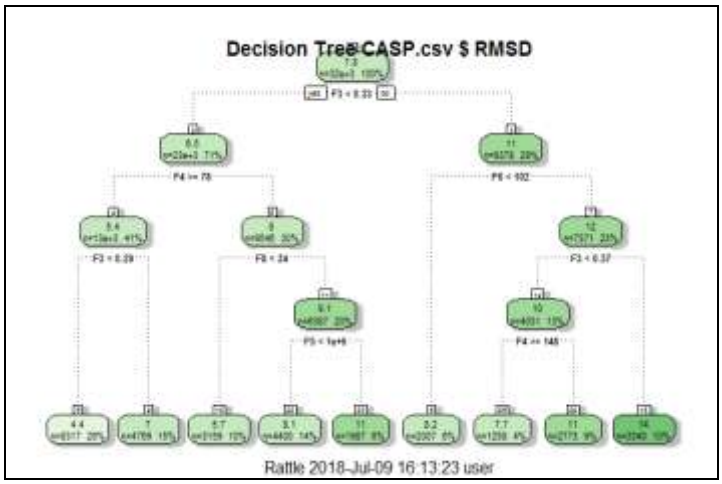

Figure 3. Regression tree for protein tertiary structure

Since the target variable of dataset is numeric and continuous, linear regression algorithm is applied for the design of regression model. The "lm()" function is used to create the model accepts formula and data. Figure 4 explains the summary of the linear regression model designed for protein tertiary structure. This output gives intercept 6.1589 and coefficients for F1 to F9. The regression equation can be derived as:

RMSD = 6.1589 + 0.0016*F1 + 0.0013*F2 + 18.5*F3 – 0.1106*F4 – 0.0235*F6 – 0.0001*F7 + 0.0149*F8 – 0.1147*F9

Different graphs plot using function plot() is shown in figure 5. The Residuals vs Fitted plot shows that linear regression model is not appropriate as regression model for protein tertiary structure dataset.

```
Call:
lm(formula = RMSD ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)])

Residuals:
     Min       1Q   Median       3Q      Max
-24.2858  -4.0797  -0.8885   3.8548  20.9556

Coefficients:
              Estimate    Std. Error t value   Pr(>|t|)
(Intercept)  6.1589824916  0.7029103674   8.762    < 2e-16 ***
F1           0.0016232085  0.0001471335  11.032    < 2e-16 ***
F2           0.0013997744  0.0001346679  10.394    < 2e-16 ***
F3          18.5020800011  1.2700911337  14.568    < 2e-16 ***
F4          -0.1106504564  0.0019682746 -56.217    < 2e-16 ***
F5          -0.0000042376  0.0000009001  -4.708 0.00000251 ***
F6          -0.0235459676  0.0020841090 -11.298    < 2e-16 ***
F7          -0.0001591041  0.0000188534  -8.439    < 2e-16 ***
F8           0.0149533659  0.0007066038  21.162    < 2e-16 ***
F9          -0.1147044795  0.0119825305  -9.573    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.162 on 32000 degrees of freedom
Multiple R-squared:  0.2695,    Adjusted R-squared:  0.2693
F-statistic:  1448 on 9 and 32000 DF,  p-value: < 2.2e-16
```

Figure 4. Summary of linear regression model

Figure 6 summarizes the neural network architecture for physicochemical properties of protein tertiary structure. The network consists of nine inputs, one hidden layer with four hidden nodes and single output i.e. RMSD. The neural net, nonlinear regression model is designed by connecting neurons to each other, feeding the numeric data through the network, combining the numbers and adjusting weight to produce a final answer [13].

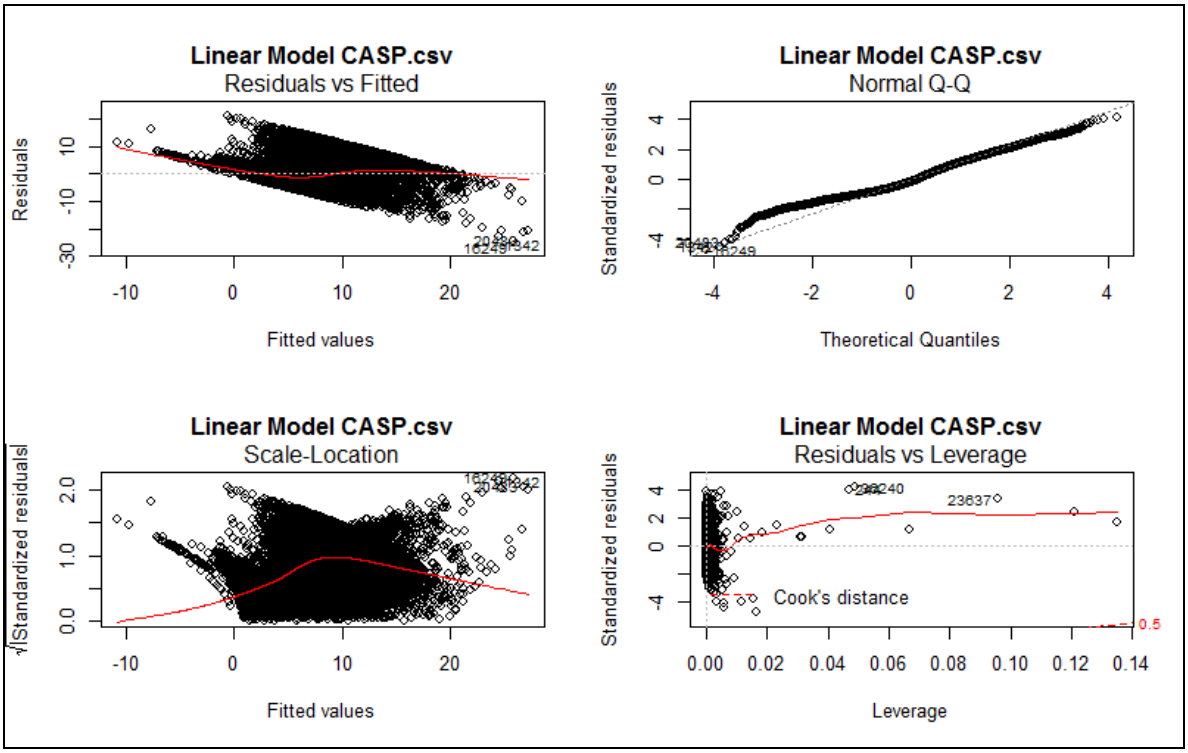

Figure 5. Linear regression model plots

. The package 'randomforest' is used for design and analysis of RF model structure [11]. The model is tuned with two parameters n_{tree} and m_{try} to get optimized structure. The summary of the model is given in figure 7. The RF model is built by using 200 decision trees and three variables are tried at each split. The mean squared residuals is 12.8435 reveals RF model is apt for regression process. Figure 8 shows measures of variable importance using RF algorithm. It depicts that physicochemical property F4 is stands at the top whereas F1 is least important.

```
A 9-4-1 network with 54 weights.
Inputs: F1, F2, F3, F4, F5, F6, F7, F8, F9.
Output: RMSD.
Sum of Squares Residuals: 852761.9528.

Neural Network build options: skip-layer connections; linear output units.

In the following table:
   b  represents the bias associated with a node
   h1 represents hidden layer node 1
   i1 represents input node 1 (i.e., input variable 1)
   o  represents the output node

Weights for node h1:
 b->h1 i1->h1 i2->h1 i3->h1 i4->h1 i5->h1 i6->h1 i7->h1 i8->h1 i9->h1
 -0.66   0.23   0.29  -0.31  -0.68  -0.36   0.27   0.23  -0.31  -0.18

Weights for node h2:
 b->h2 i1->h2 i2->h2 i3->h2 i4->h2 i5->h2 i6->h2 i7->h2 i8->h2 i9->h2
  0.31  -0.02   0.29  -0.50   0.39   0.25  -0.16  -0.55  -0.52   0.25

Weights for node h3:
 b->h3 i1->h3 i2->h3 i3->h3 i4->h3 i5->h3 i6->h3 i7->h3 i8->h3 i9->h3
 -0.65  -0.15  -0.03  -0.20   0.30  -0.16  -0.04   0.49   0.56   0.44

Weights for node h4:
 b->h4 i1->h4 i2->h4 i3->h4 i4->h4 i5->h4 i6->h4 i7->h4 i8->h4 i9->h4
  0.41   0.51   0.38   0.22   0.47  -0.41   0.15  -0.22   0.46  -0.08

Weights for node o:
 b->o h1->o h2->o h3->o h4->o i1->o i2->o i3->o i4->o i5->o i6->o i7->o i8->o
 3.15  0.33  3.01  0.56  0.59  0.00  0.00 18.50 -0.11  0.00 -0.02  0.00  0.01
i9->o
-0.11
```

Figure 6. Neural network model for protein tertiary structure

```
Number of observations used to build the model: 32010
Missing value imputation is active.

Call:
 randomForest(formula = RMSD ~ .,
              data = crs$dataset[crs$sample, c(crs$input, crs$target)],
              ntree = 200, mtry = 3, importance = TRUE, replace = FALSE, na.action = randomForest::na.roughfix)

               Type of random forest: regression
                     Number of trees: 200
No. of variables tried at each split: 3

          Mean of squared residuals: 12.84357
                    % Var explained: 65.74

Variable Importance
===================

   %IncMSE IncNodePurity
F8   94.02      86199.77
F4   76.89     121589.47
F3   67.93     119011.35
F6   53.99      71784.37
F7   49.51      66879.62
F5   47.09      55713.64
F1   43.12      52232.60
F2   38.55      73627.19
F9   38.10      58508.39
```

Figure 7. Summary of RF regression model

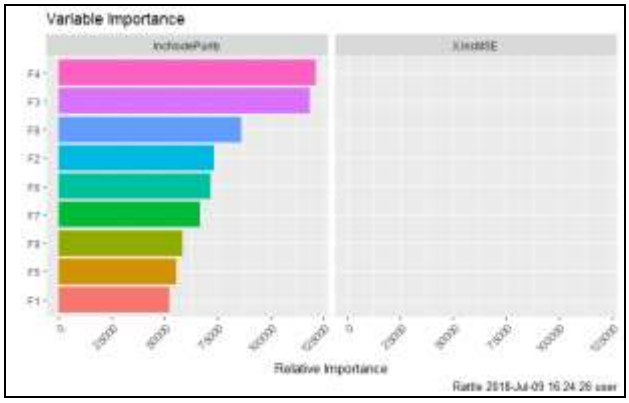

Figure 8. Variable importance measure

Performance of the models is evaluated with reference Predicted Versus Observed plot for test dataset. It plots predicted values against observed values as shown in figure 9. The plot also gives R-squared value. A better model is one with value of R-squared closer to 1. Table 1 briefs R-Squared value for different prediction algorithms applied on protein tertiary structure dataset. Performance accuracy reveals Random Forest model is appropriate as regression model for Physicochemical Properties of Protein Tertiary Structure as compared to others.

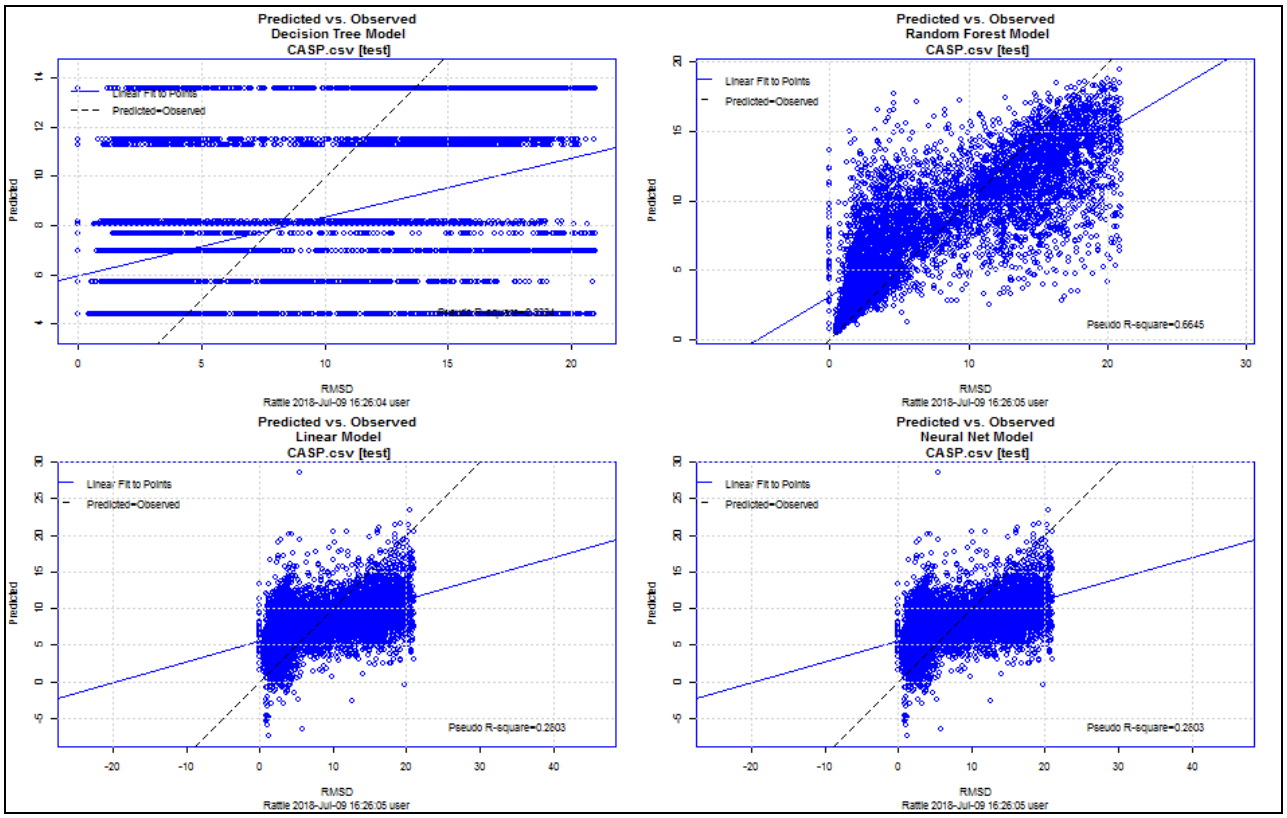

Figure 9. Predicted Versus Observed plot of regression models on test dataset

Table 1. Performance accuracy of regression models

Regression Model	R-Squared
Decision Tree	0.2224
Random Forest	0.6645
Linear Model	0.2803
Neural Network	0.2803

IV. CONCLUSION

This paper exhibits performance evaluation of various regression algorithms for the prediction of protein tertiary structure by modelling physicochemical properties. The physical and chemical properties of protein determine quality of protein structures. Performance accuracy reveals Random Forest model predicts RMSD value with less error on test data as compared to other models. The dataset for the present study is retrieved from UCI machine learning repository. The result reveals that the Random Forest model outperforms the rest of the model in prediction of protein tertiary structure. The measures of variable importance using RF algorithm reveals that physicochemical property F4 is stands at the top whereas F1 is least important.

REFERENCES

[1] Dua, D., Karra, T. E., "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]". Irvine, CA: University of California, School of Information and Computer Science, 2017.

[2] Cozzone, A. J., "Proteins: Fundamental Chemical Properties, Encyclopedia of Life Sciences", Macmillan Publishers Ltd, 2002.

[3] Mishra, S., Pathak, T., Ahirwar, A., "Classification of Protein Structure (RMSD <= 6A˚) using Physicochemical Properties", International Journal of Bio-Science and Bio-Technology, Vol 7, No. 6, pp.141-150, 2015.

[4] Gromiha, M. M., "A Statistical Model for Predicting Protein Folding Rates from Amino Acid Sequence with Structural Class", Information, J. Chem. Inf. Model, Vol 45, pp. 494-501, 2005.

[5] Pathak, Y., Rana, P. S., Singh, P.K., "Protein structure prediction (RMSD < = 5 Å) using machine learning models", Int. J. Data Mining and Bioinformatics, Vol 4, No. 1, pp. 71-85, 2016.

[6] Jani, A. M., Chandpa, K. R., "Protein Tertiary Structure Classification based on its Physicochemical property using Neural Network and KPCA-SVM: A Comparative Study", Int. Jour. of App. Sc. and En, Vol 3, No. 1, pp. 1-11, 2015.

[7] Mishra, S., Ahirwar, A., "Comparative Study of Machine Learning Models in Protein Structure Prediction", International Journal of Computer Science and Information Technologies, Vol 6, No. 6, pp. 5398-5404, 2015.

[8] Iraji, M. S., Ameri, H., "RMSD Protein Tertiary Structure Prediction with Soft Computing", I.J. Mathematical Sciences and Computing, Vol 2, pp. 24-33, 2016.

[9] Kamath, R.S., Kamat, R.K., "Modeling Mice Down Syndrome through Protein Expression: A Decision Tree based Approach",

Research Journal of Pharmaceutical, Biological and Chemical Sciences, Vol 7, No. 4, pp. 1193-1199, 2016.

[10] Kamath, R.S., Kamat, R.K., "Educational Data Mining with R and Rattle", River Publishers Series in Information Science and Technology, Netherland, 2016.

[11] Kamath, R.S., Dongale, T.D., Pawar, P., Kamat, R.K., "Random Forest Modeling for Mice Down Syndrome through Protein Expression: A Supervised Learning Approach", Research Journal of Pharmaceutical, Biological and Chemical Sciences, Vol 7, No. 4, pp. 830-836, 2016.

[12] Kamath, R.S., Kamat, R.K., "Modelling Fetal Morphologic Patterns Through Cardiotocography Data: Decision Tree Based Approach", Journal of Pharmacy Research, Vol 12, No. 1, pp. 9-12, 2017.

[13] Kamath, R.S., "Design and Development of Soft Computing Model for Teaching Staff Performance Evaluation", International Journal of Engineering Sciences & Research Technology, Vol 3, No. 4, pp. 3088-3094, 2014.

AUTHOR'S BIOGRAPHIES

Dr. Mrs. R.S. Kamath is Associate Professor in the Department of Computer Studies, Chhatrapati Shahu Institute of Business Education and Research, Kolhapur, India. She obtained her Bachelors and Masters in Computer Science from Mangalore University. She received her Ph.D. in Computer Science specialized in Computer Based Visualization from Shivaji University and completed the same in 2011. Dr. Kamath has to her credit 35 research papers published in reputed national and international journals and presented 14 papers in national conferences. She has completed two minor research funded by UGC. She is the author of four books and the chapter entitled "Cost Effective 3D Stereo Visualization for Creative Learning – Virtual Reality in Education" is published in Encyclopedia of Information Science and Technology 4th Edition by IGI Global Publishing. She has authored the book "Educational Data Mining with R and Rattle published by River Publishers Series in Information Science and Technology, Netherland. Her areas of research interests are Artificial Intelligence, Machine Learning, Virtual Reality and Soft Computing. She has immense skill of around fourteen years in teaching and research.

Dr. R.K. Kamat holds the position of Professor in the Department of Electronics and heads the Department of Computer Science of Shivaji University, Kolhapur. He is Director of Internal Quality Assurance (IQAC) of the Shivaji University, Kolhapur. He obtained his Bachalors, Masters and M.Phil in Electronics from the Shivaji University, Kolhapur. Dr. Kamat gained his Ph.D. specialized in Smart Sensors from Goa University, Goa. Professor Kamat has published over 70 plus papers in International journals of repute and
presented equal number of papers at National and International Conferences. He has published 10 books though reputed publishing house such as Springer UK. He has published scholarly literature on the quality issues in higher education. Five students have been awarded Ph.D. under his guidance and 11 more are working for their Doctorate. He has been involved with various initiatives of the apex organization at international level such as IEEE USA and Engineering Education group of Central Quinsland Australia. Through these organizations he is playing key role in spreading the scholastic culture by organizing conferences at various international destinations. Through this network he has visited various countries. Dr. Kamat is a recipient of the Young Scientist Award under the fast track scheme of the Department of Science and Technology (DST) of Government of India.